ChinaXiv合作期刊

第 38 卷第 7 期
2018 年 4 月

生 态 学 报
ACTA ECOLOGICA SINICA

Vol.38,No.7
Apr.,2018

DOI: 10.5846/stxb201705030814

陈妍,宋豫秦,王伟.基于随机森林回归的草场植被盖度反演模型研究——以新疆阿勒泰地区布尔津县为例.生态学报,2018,38(7):2384-2394.

Chen Y, Song Y Q, Wang W. Grassland vegetation cover inversion model based on random forest regression: A case study in Burqin County, Altay, Xinjiang Uygur Autonomous Region. Acta Ecologica Sinica,2018,38(7):2384-2394.

基于随机森林回归的草场植被盖度反演模型研究
——以新疆阿勒泰地区布尔津县为例

陈 妍[1], 宋豫秦[1,*], 王 伟[2]

1 北京大学环境科学与工程学院, 北京 100871

2 中国环境科学研究院生物多样性研究中心,北京 100012

摘要:作为草地资源大国,我国正面临严峻的草场退化形势。掌握草场植被盖度的历史演变趋势,是草场退化驱动力识别及风险评估的基础。目前已有研究多以参数回归方法估算植被盖度,但并未充分考虑其苛刻的使用条件。利用 Landsat 系列卫星遥感影像及地面植被盖度监测资料建立非参数回归——随机森林回归模型,并与传统线性回归方法进行比较,在此基础上应用随机森林回归模型估算近 10 年来布尔津县草场植被盖度的变化趋势,并对结果的不确定性进行分析。结果显示:传统的线性回归方法很难满足其基本的统计学假设条件,而随机森林模型不但无需进行假设条件检验,而且预测的准确性也优于以往普遍应用的线性模型。基于 Landsat ETM+标准数据得到的反演结果较之 TM 和 OLI 数据普遍偏小,地表反射率数据虽然可以大幅降低传感器不同对反演结果所造成的影响,但结果仍存在约±10%的不确定性。涉及的草场类型众多,为了提高反演精度,后续研究需要分别计算其植被指数,并尽量减低传感器差异带来的不确定性。

关键词:植被盖度;植被指数;随机森林;遥感影像

Grassland vegetation cover inversion model based on random forest regression: A case study in Burqin County, Altay, Xinjiang Uygur Autonomous Region

CHEN Yan[1], SONG Yuqin[1,*], WANG Wei[2]

1 *College of Environmental Sciences and Engineering, Peking University, Beijing* 100871, *China*

2 *Biodiversity Research Center, Chinese Research Academy of Environmental Sciences, Beijing* 100012, *China*

Abstract: As a large country with extensive grassland resources, China is facing severe grassland degradation. Studying trends in grassland vegetation cover change provides a basis for identifying the driving forces of grassland degradation and associated risk assessment. In previous studies, parametric regression models have typically been applied to estimate vegetation cover. However, the harsh assumptions of parametric regression have always been neglected. In the current study, vegetation cover monitoring data and vegetation indices (NDVI, SAVI, MSAVI, EVI), extracted from Landsat remote sensing images, were used to build random forest regressions, which are non-parametric models. These models were subsequently compared with traditional linear regression models. To build and test these models, 205 samples were collected from 2010 to 2015 (data from 2012 were not included) in alpine meadow, mountain meadow, lower-flat meadow, temperate meadow steppe, desert steppe, steppe desert, and desert in Burqin County, Xinjiang Uygur Autonomous Region. Among these samples, 150 samples were used to build models, and the remainder was used as testing data. Two sets of Landsat remote sensing images, Level 1 Standard Product and Surface Reflectance Product, were applied separately, and both

基金项目:国家重点研发计划(2016YFC0503300)

收稿日期:2017-05-03; **网络出版日期**:2018-01-04

* 通讯作者 Corresponding author.E-mail: yqsong@pku.edu.cn

chinaXiv:201804.01739v1

ChinaXiv合作期刊

included TM data for 2011—2012 and OLI data for 2013—2015. In total, two random forest models and 23 linear models were built. The results indicated that the predictive ability of the random forest models was clearly stronger than that of most of the linear models. Moreover, it was not necessary to test the assumptions for the random forest models, whereas none of the linear models' assumptions in this study could be satisfied perfectly. In the case study, random forest regression was applied to estimate the trend in grassland vegetation cover change in the last decade in the study area based on 663 sampling points. Among these, data for 2005—2009 were based on Landsat ETM+, data for 2010—2011 were based on Landsat TM, and data for 2013—2015 were based on Landsat OLI. It was clear that sensor differences would have a certain influence on the inversion results. Therefore, we also simultaneously built a random forest model for MODIS-EVI data, as this would not be affected by sensor differences, and the results calculated using MODIS data were considered to be a standard. For Level 1 standard data, the results based on Landsat ETM+ were significantly smaller than the results based on MODIS data. For surface reflectance data, the influence of different sensors on the results could be markedly reduced. Finally, to quantify the uncertainty of vegetation cover change trend based on surface reflectance data, we used a random forest model to verify vegetation cover extracted from different sensors during the same period, and found that the uncertainty was between −10% and 11%. In conclusion, random forest regression is assumed to be a better model to inverse vegetation cover than linear models. For its application in time series studies, Landsat surface reflectance production could significantly reduce the influence of sensor difference, although the uncertainty was still approximately ±10%. In the current study, we assessed many grassland types, and to improve the accuracy of prediction, vegetation indexes should be calculated separately in further studies. In addition, efforts should be made to reduce the uncertainty associated with the data from different types of sensor.

Key Words: vegetation cover; vegetation index; random forest; remote sensing images

我国是草地资源大国,但多年来的超载过牧、荒地过度开垦和滥挖乱采致使草场植被覆盖度大幅下降,水土流失和沙漠化日趋严重,牧区经济发展及草地生态系统的健康因此受到了严重威胁[1]。研究草场变化规律是揭示草场退化驱动力以及风险评估的基础,在缺乏长期监测数据的情况下,借助卫星遥感影像不失为有效的方法。当前可用的遥感数据一般为气象卫星 NOAA 的 1km 分辨率 AVHRR 数据[2-3],terra 和 aqua 卫星的 250m 分辨率 MODIS 数据[4-5]和陆地卫星 Landsat 系列数据。早期研究多采用分辨率较低的 AVHRR 数据,MODIS 数据虽具有较高的完整性,但仍难以满足长时间序列研究的需求。Landsat 系列卫星自 1975 年开始已有 7 颗卫星相继运行,且其 30m 的分辨率相较于 AVHRR 和 MODIS 来说优势明显,因而被目前不少研究者所关注。

利用遥感影像研究草场变化趋势大致可分为两种思路,其一是通过监督或非监督分类,分析草场分布或面积的变化[6-8],然而面积增减只能体现宏观变化,难以反映某一具体区域在时间序列上的退化或恢复情况;其二是利用光谱信息计算植被指数,通过植被指数变化直接表征生态系统的变化趋势[9-11],相对而言,这类方法的主要优势是将研究对象精确至像元尺度。然而干旱半干旱地区植被盖度低,植被光谱信息弱,植被指数往往非常小,在这种情况下用其变化情况说明草场生态的改善或恶化往往意义不大。而利用植被指数反演植被盖度不失为一种可行的替代思路。植被盖度计算主要可通过线性光谱混合模型[12-14]和回归模型两种方法实现。前者一般应用于缺乏地面监测数据的情况之下,其核心是获得纯像元下的地物光谱值。但在实际应用中,特别是植被盖度较低的草地生态系统中,典型光谱值很难获得,且计算误差较大[15]。因此,在具备地面监测数据的情况下,一般通过植被指数与植被盖度直接的相关关系建立参数模型,反演植被盖度,其中线性回归模型应用最为广泛[16-17]。然而以往的研究多未充分考虑参数回归苛刻的假设条件,以及多元回归对变量间非共线性的要求,这无疑会降低反演模型的可靠性。解决此问题的途径之一是寻求预测效果达到甚至超越参数模型的非参数方法。作为目前预测效果最好的非参数回归模型之一,随机森林模型与参数回归等方法相

chinaXiv:201804.01739v1
www.ecologica.cn

ChinaXiv合作期刊

比,无需对变量的正态性和独立性等假设条件进行检验,同时也不需要考虑多变量的共线问题[18],且运算高效、结果准确,在环境以及生态学等领域都有着广泛应用[19-22]。在草场研究方面,随机森林虽曾用于植被分类[23],但作为回归模型的应用则几属空白。

数据源是保证反演模型可靠性的前提,长时间序列研究会不可避免地使用到来自不同传感器的光谱信息,而对于 Landsat 系列卫星标准数据产品而言,同一时间段内从不同传感器所提取的光谱信息会存在一定差异,需要对非同源数据进行校正以及标准化处理[24-25]。但目前草场方面已有的研究既没有定量分析传感器差异对反演结果造成的影响,也没有计算校正和标准化处理对反演结果的改进效果。

鉴此,本文利用 Landsat 系列卫星 TM,ETM+以及 OLI 的影像和植被盖度地面监测数据建立基于随机森林回归的草场植被盖度反演模型,并将反演结果与线性回归结果进行比对,探讨其优越性。继之将其应用于基于多源数据的草场植被盖度变化趋势分析,并探讨该方法的不确定性,以期为干旱半干旱地区长时间序列的草场变化研究提供方法支持。

1 研究区域

布尔津县位于新疆西北部阿勒泰地区,属阿尔泰山脉西南麓,介于 86°25′—88°6′E,47°22′—49°11′N 之间,总面积 10369km^2。境内景观异质性强,北部山区最高蜂海拔 4374m,中部为低山丘陵、河谷地带,南部为低平戈壁滩,海拔最低处仅为 436m(图 1)。全县生态系统多样,北部山区兼具水源涵养与生物多样性保护的功能,南部荒漠生态系统肩负防风固沙的生态功能。草地资源丰富,各类草地类型均有分布,农业以畜牧为主,放牧方式仍属传统游牧。当前北部林区载畜量快速增加,林牧矛盾突出,南部以荒漠植被为主,生态环境脆弱,生态保护形势更加严峻。丰富而多样的草场分布特征,不仅可使草场植被盖度反演模型为分析当地草场历史变化趋势乃至制定草场保护和畜牧业生产政策奠定基础,同时也可为其他地区的各类草场研究提供一定借鉴。

2 数据获取与预处理

为了建立基于随机森林回归的草场植被盖度反演模型,本文所用之数据主要包括基于地面调查的草场盖度数据和卫星遥感影像。

植被盖度地面调查数据来源于布尔津县畜牧兽医局草原站提供的 2010、2011、2013、2014、2015 年 6—9 月的草本、半灌木及矮小灌木草原样方调查表。调查样地涵盖全县 8 个典型草场类型,即低平地草甸、高寒草甸、山地草甸、温性草甸草原、温性草原化荒漠、温性荒漠以及温性荒漠草原。依据典型性原则,样方在能够代表整个样地草原植被、地形及土壤等特征的区域随机布设,监测样地共计 205 个,分布及类型如图 1 所示,每个样地设置 3 个样方,样方大小为 1m×1m,监测结果用 3 个样方的平均值表示。部分监测点位虽然超越了县界,但并未超越本文所用的遥感影像范围,因此本文将此类监测点也纳入分析范围。植被盖度是由目测法测定的样方内植物地上部分垂直投影面积占样方面积的比率,监测频次为每年一次。

遥感数据为下载自美国地质调查局(U.S. Geological Survey)网站(http://glovis.usgs.gov/)的 Landsat 系列卫星遥感影像,整个县域需要两景影像覆盖,行带号分别为 144/26、144/27,影像时间及传感器信息如表 1 所示。

本文所使用的每一景影像都包含有目前此类研究通用的一级标准数据产品(简称 L1 数据)以及经过校正和标准化处理的地表反射率数据产品(简称 SR 数据),用以定量研究传感器差异对反演结果的影响。

本文的研究对象为草场,因此需要对影像进行监督分类,提取出草场范围。首先,利用 ENVI 5.1 软件对 2015 年 Landsat8 OLI 的两景 L1 影像(144/27、144/27)分别进行剪裁、辐射定标、融合以及 FLAASH 大气校正等预处理。其次初步将影像定义为林地、草地、荒地、水体、冰雪区五大类型,分别划定训练样本,通过最大似然法进行分类。经过分类后处理,以 Google Earth Pro. 的高分辨率影像为基础,划定验证样本,通过混肴矩阵

chinaXiv:201804.01739v1

ChinaXiv合作期刊

chinaXiv:201804.01739v1

计算分类精度。两景影像(144/27、144/27)的总体分类精度分别为 97.7705% 和 98.8749%,Kappa 系数分别为 0.9692 和 0.9642。合并草地、荒地两种类型的区域,通过目视解译去除其中的耕地和建设用地区域,得到布尔津县草场分布范围(图 2)。

3　研究方法

3.1　随机森林模型的建立与验证

3.1.1　植被指数选择

归一化植被指数(NDVI)是草场退化研究中使用最为广泛的植被指数[26-28],本文首先选择其作为模型的解释变量之一。由于研究区南部大片区域为植被盖度较低的温性草原化荒漠、温性荒漠以及温性荒漠草原,仅使用 NDVI 作为解释变量,反演效果可能并不能达到预期,必须考虑土壤背景的影响[29],因此进一步引入可用于修正土壤背景敏感性以及气溶胶散射影响三个植被指数——土壤调节植被指数(SAVI)、修正土壤调节植被指数(MSAVI)、增强植被指数(EVI)。上述植被指数的计算方法如式 1—式 4 所示,其中参数的选取依据了 USGS 发布的植被指数产品相关指南[30]。

$$NDVI=\frac{NIR-RED}{NIR+RED} \tag{1}$$

$$SAVI=\frac{NIR-RED}{NIR+RED+0.5}\times1.5 \tag{2}$$

$$MSAVI=\left(2\times NIR+1-\sqrt{(2\times NIR+1)^2-8\times(NIR-RED)}\right)\times0.5 \tag{3}$$

$$EVI=2.5\times\frac{NIR-RED}{NIR+6\times RED-7.5\times BLUE+1} \tag{4}$$

3.1.2　模型建立

图 1 中监测点位的植被盖度数据分别来自 2010、2011、2013、2014、2015 年 6—9 月,经统计分析,在 6—9 月期间,植被盖度并无显著区别,因此在对应年份的 6—9 月中选择云量最少的影像加以利用,即利用 2010—2011 年 8 月的 TM 影像和 2013—2015 年 7 月的 OLI 影像计算相应点位的 4 个植被指数。

利用 R 软件在 205 组数数据(植被盖度及其对应植被指数)中随机抽取 150 组作为训练数据用于模型建立。

表 1　遥感影像信息列表

Table 1　Information of satellite images

影像时间 Image dates	传感器信息 Information of sensors	
2016-08	Landsat8 OLI	Landsat7 ETM+
2015-07	Landsat8 OLI	Landsat7 ETM+
2014-07	Landsat8 OLI	
2013-07	Landsat8 OLI	
2011-08	Landsat4-5 TM	
2010-08	Landsat4-5 TM	
2009-08	Landsat4-5 TM	Landsat7 ETM+
2008-08	Landsat7 ETM+	
2007-08	Landsat7 ETM+	
2006-07	Landsat7 ETM+	
2005-08	Landsat7 ETM+	

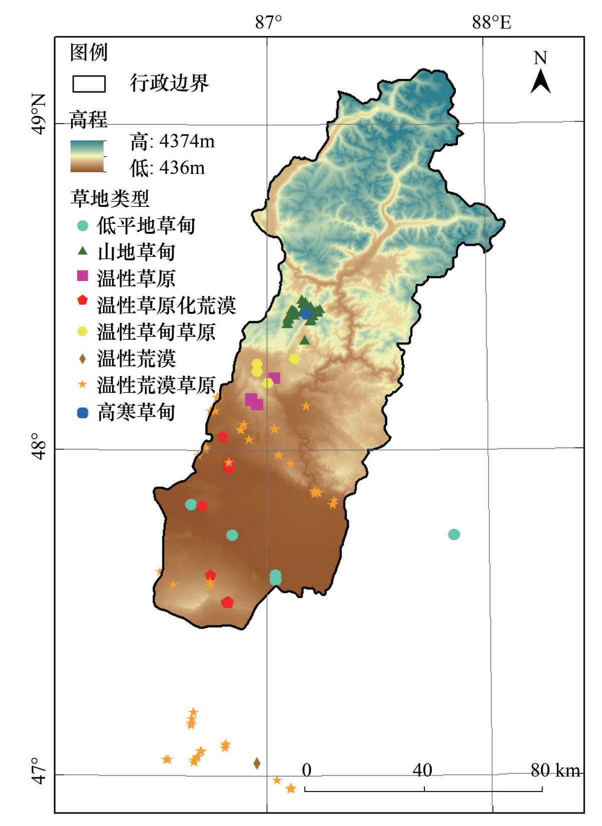


图 1　研究区地形及监测点位分布图

Fig.1　Elevation and monitoring points' distribution of study area

随机森林是由多棵相互没有关联的决策树组成的集成决策树,用{$h(X,\Theta k)$, $k=1$, …}表示,其中 X 为输入向量,{Θk}为独立同分布随机向量[31]。与经典回顾分析类似,随机森林可以解释多个自变量(X_1、X_2…X_k)对因变量 Y 的作用。此处将植被盖度的实地监测值作为因变量 Y,NDVI、SAVI、MSAVI 和 EVI 作为自变

www.ecologica.cn

ChinaXiv合作期刊

量。随机森林模型的建立通过调用 R 语言中“randomForests”程序包[32]来实现。该方法首先完成两个随机采样过程,即通过自助法重采样技术有放回的在150组训练数据中重复随机抽取150个训练样本,未被抽取到的数据被称为“袋外”(out of bag)数据。再从NDVI、SAVI、MSAVI 和 EVI 的4个变量中随机选择若干个变量(以 m_{try} 表示)建立决策树,模型中 m_{try} 的省缺值一般为总变量的1/3,本文有4个变量,则 m_{try} 值为1。最后重复 n 次上述过程,生成由 n 棵决策树组成的随机森林,n 值越大预测越好。随着 n 值的增大,袋外数据误差在显著降低后会基本保持稳定。为了节省计算时间,n 值保证袋外数据误差稳定即可,本文 n 值取150。

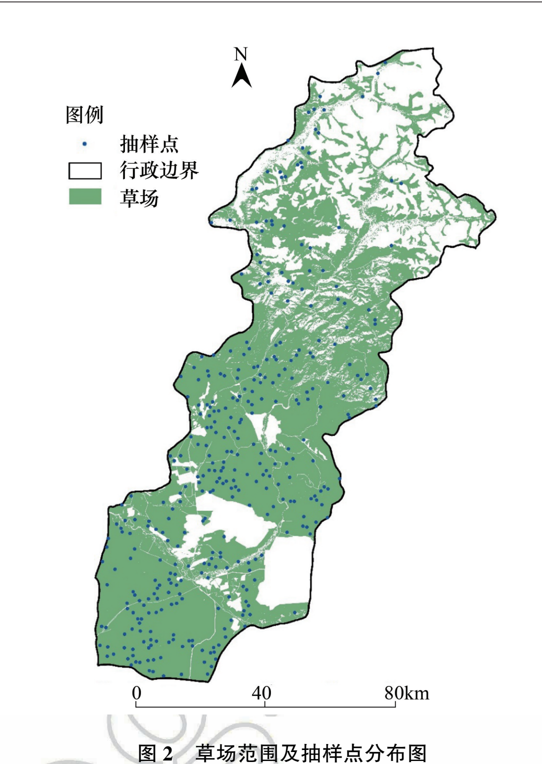


图2 草场范围及抽样点分布图

Fig.2 Distribution of grassland and sampling points in study area

模型输出所有决策树计算结果的均值,并通过计算“解释方差百分比”来评定模型预测能力。对于各个自变量对因变量的影响程度,用方差增量以及节点纯度增量两个指标来定性表征。前者指将某一变量替换成随机变量后对预测结果造成的影响,若用于替换的随机变量显著改变了方差,则认为原变量重要性很高;后者从同质性增加的角度去表征变量的重要性[33]。随机森林模型的具体算法请参照 Biau 等[34]的文章。

为了将随机森林回归方法与传统的线性回归进行比较,本文采用同样的训练数据同时建立了植被盖度与NDVI、SAVI、MSAVI 和 EVI 的一元线性回归及多元线性回归模型。

3.1.3 模型验证

利用205组数据经过随机抽样剩余的55组数据用于模型验证。应用前文建立的随机森林模型和各个线性回归模型分别计算55个点位的植被盖度预测值,并将实测值与预测值相减,根据差值分布情况比较模型的预测能力。此外,通过绘制残差散点图和正态 Q-Q 图来检验线性回归模型方差齐性以及残差的正态性,并通过 Kappa 系数来判断变量的共线性。

3.2 模型的应用及其不确定性分析

3.2.1 模型应用

利用 ArcMap 10 软件在研究区域内随机生成1000个相互间隔大于1km 的点位,去除草场范围之外的点位,得到草场抽样点。为了研究模型在基于多源数据的植被盖度反演领域的应用,本研究以2005—2008年ETM+影像、2009—2011年 TM 影像、2013—2015年 OLI 影像为基础,分析2005—2015年间(除2012年),植被盖度的变化趋势。

首先,剔除研究时段内落在云层和云影中的草场抽样点,最终得到的356个点位用于本小节的抽样分析(图2)。其次,分别利用上述影像的 L1 和 SR 数据计算各个点位的 NDVI、SAVI、MSAVI 和 EVI 值,将其输入3.1中建立的随机森林模型,分别得到监测点位基于 L1 数据和 SR 数据的植被盖度值。最后,将每年356个监测点位的植被盖度预测值求平均,绘制出基于 L1 数据和 SR 数据的植被盖度变化趋势图。

3.2.2 不确定性分析

由于2005—2015年间遥感影像源自3个不同的传感器,而传感器的差异会给反演结果带来一定影响,因此本研究进一步应用下载自美国地质调查局地球资源观察和科学中心(Earth Resources Observation and Science Center)的 MODIS—MOD13Q1 V005 数据集建立随机森林模型,反演2005—2015年的植被盖度。由于

chinaXiv:201804.01739v1

ChinaXiv合作期刊

MODIS 数据不存在传感器差异问题,因此可将基于此数据反演出的植被盖度变化趋势作为参考,将 3.2.1 中得到的基于 L1 和 SR 数据的趋势曲线与之相比较,变化趋势越接近基于 MODIS 数据的反演结果,则说明该数据中不同传感器给反演结果造成的误差越小。

前文提到 SR 数据是 L1 数据经过校正和标准化处理的地表反射率数据产品,可在一定程度上克服传感器的差异对结果带来的影响。为了定量分析 SR 数据对反演结果的改进效果及不确定性,本研究进一步比对同一时期内,基于不同传感器的反演结果之间存在的差异。

研究区范围草场生长期内,同时具有 TM 和 ETM+影像的时段是 2009 年 8 月,同时具有 ETM+和 OLI 影像的时段是 2015 年 7 月和 2016 年 8 月。在草场抽样点中去除 09、15、16 年 3 年内云层、云影中的点位以及 ETM+数据中坏条带中的点位后,剩余 663 个点位用于不确定分析。利用随机森林模型分别计算同一时段基于不同传感器的反演结果并相减,分别绘制基于 L1 数据和 SR 数据的 TM—ETM+ 、OLI—ETM+的盖度反演结果差异概率密度图。

4　结果与讨论

4.1　拟合结果

随机森林模型的计算结果显示,基于 L1 数据和 SR 数据而建立的随机森林回归模型输出的解释方差百分比分别为 69.82%和 72.42%,因此后者的预测效果优于前者;重要性方面,各个植被指数并未表现出明显的差别(表 2)。

表 2　随机森林回归模型变量重要性

Table 2　The variables' importance in random forest models

变量 Variables	一级标准数据产品 Level1 standard data product		地表反射率数据产品 Surface reflectance data products	
	方差增量 Increase in mean squared error	节点纯度增量 Increase in node purity	方差增量 Increase in mean squared error	节点纯度增量 Increase in node purity
NDVI	12.927	35899.58	9.119	27067.63
SAVI	10.777	31475.49	9.080	29956.79
MSAVI	11.111	38354.48	7.008	23012.77
EVI	—	—	7.837	26277.86

由于 L1 数据未经过标准化处理,TM 和 OLI 数据计算的 EVI 值不在同一数量级,无法用于模型的建立;NDVI:归一化植被指数,Normalized Difference Vegetation Index;SAVI:土壤调节植被指数,Soil Adjusted Vegetation Index;MSAVI:修正土壤调节植被指数,Modified Soil Adjusted Vegetation Index;EVI:增强植被指数,Enhanced Vegetation Index

本研究一共建立线性回归模型 23 个,其中 P 值显著的模型及相应参数如表 3 所示。由表 3 可知,对于一元线性回归模型来说 L1 数据和 SR 数据的 R^2差别不大;对于多元回归来说所有的 L1 数据都无法得到显著结果;从 R^2的数值上来看 SR 数据的多元回归的效果要明显优于一元回归。

4.2　模型比较

为了进一步分析上述模型的预测能力,本文将 55 个验证点位的监测值和预测值作差,结果的分布情况如图 3 所示。首先,虽然 L1 和 SR 数据的一元回归效果在表 3 中无法通过 R^2判断,但通过比较红色和绿色箱线图可明显看出,SR 数据实测值和预测值的差异明显小于 L1 数据;第二,随机森林回归模型的预测能力略优于多元回归模型,且这两者明显优于一元线性回归;第三,所有线性回归的结果较真实值来说普遍偏小,但随机森林的预测偏差相对均衡。

在本文所建立的所有模型中,预测效果最好的是基于 L1 和 SR 数据的随机森林模型,以及由变量 SAVI-MSAVI 建立的 SR 数据二元回归模型。

chinaXiv:201804.01739v1

ChinaXiv合作期刊

chinaXiv:201804.01739v1

表 3 线性回归模型结果

Table 3 The results of linear models

模型 Models		一级标准数据产品 Level1 standard data product			地表反射率数据产品 Surface reflectance data products		
		值 Value	P	R^2	值 Value	P	R^2
EVI	常量 Constant		—		25.13	<0.001	0.65
	系数 Parameter				0.01	<0.001	
NDVI	常量 Constant	32.49	<0.001	0.68	20.54	<0.001	0.73
	系数 Parameter	124.42	<0.001		0.01	<0.001	
SAVI	常量 Constant	32.49	<0.001	0.68	22.46	<0.001	0.69
	系数 Parameter	83.07	<0.001		0.01	<0.001	
MSAVI	常量 Constant	32.1	<0.001	0.66	26.55	<0.001	0.64
	系数 Parameter	85.69	<0.001		0.01	<0.001	
SAVI+MSAVI	常量 Constant		不显著		9.63	<0.001	0.77
	系数 1 Parameter1				0.08	<0.001	
	系数 2 Parameter2				−0.06	<0.001	
SAVI+EVI	常量 Constant		不显著		17.94	<0.001	0.73
	系数 1 Parameter1				0.05	<0.001	
	系数 2 Parameter2				−0.04	<0.001	
NDVI+SAVI+EVI	常量 Constant		不显著		17.93	<0.001	0.74
	系数 1 Parameter1				0.01	<0.05	
	系数 2 Parameter2				0.03	<0.05	
	系数 3 Parameter3				−0.02	<0.05	

由于 L1 标准化处理,TM 和 OLI 数据计算的 EVI 值不在同一数量级,无法用于模型的建立

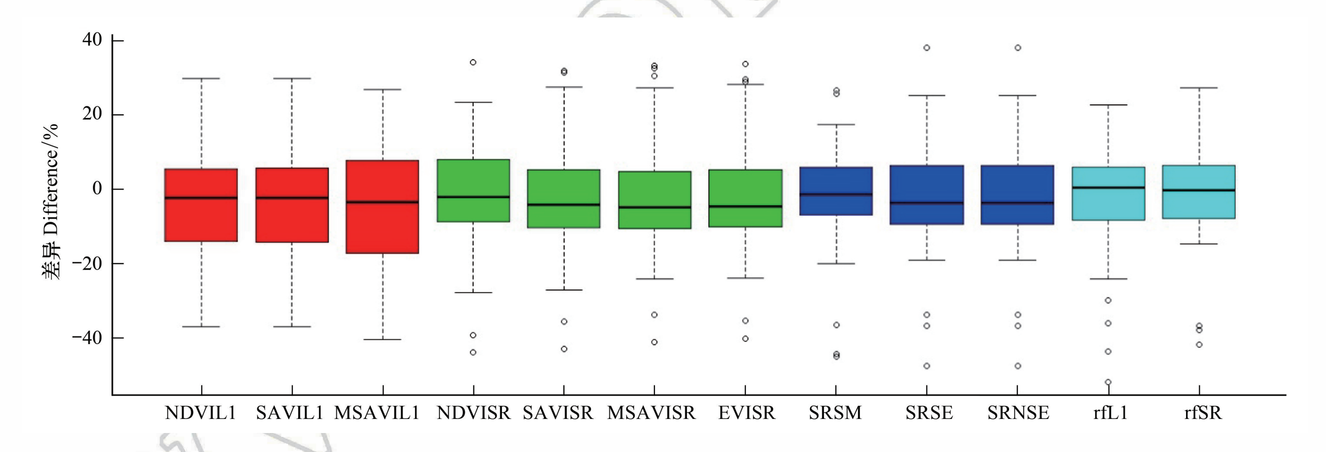


图 3 植被盖度的实测值与预测值差异

Fig.3 The difference between monitoring values and predicted values

NDVIL1:基于 L1 数据 NDVI 指数的线性回归模型,Linear model based on NDVI from level 1 standard product;SAVIL1:基于 L1 数据 SAVI 指数的线性回归模型,Linear model based on SAVI from level 1 standard product; MSAVIL1:基于 L1 数据 MSAVI 指数的线性回归模型,Linear model based on MSAVI from level 1 standard product;NDVISR:基于 SR 数据 NDVI 指数的线性回归模型,Linear model based on NDVI from surface reflectance product;SAVISR:基于 SR 数据 SAVI 指数的线性回归模型,Linear model based on SAVI from surface reflectance product;MSAVISR:基于 SR 数据 MSAVI 指数的线性回归模型,Linear model based on MSAVI from surface reflectance product;EVISR:基于 SR 数据 EVI 指数的线性回归模型,Linear model based on EVI from surface reflectance product;SRSM:基于 SR 数据 SAVI 及 MSAVI 指数的线性回归模型,Linear model based on SAVI and MSAVI from surface reflectance product;SRSE:基于 SR 数据 SAVI 及 EVI 指数的线性回归模型,Linear model based on SAVI and EVI from surface reflectance product;SRNSE:基于 SR 数据 NDVI、SAVI 及 EVI 指数的线性回归模型,Linear model based on NDVI,SAVI and EVI from surface reflectance product;rfL1:基于 L1 数据的随机森林回归模型,Random forest model based on level 1 standard product;rfSR:基于 SR 数据的随机森林回归模型,Random forest model based on surface reflectance product

ChinaXiv合作期刊

图 4 为 SRSM 方差齐性和残差正态性检验结果。理想状态下,残差-拟合值关系图上的点应随机分布,而正态 QQ 图中的点应分布于虚线之上,但 SRSM 模型并不能满足其基本假设条件,特别是残差的正态分布性。不仅如此,本研究所有其他的线性回归模型也都不能很好的满足其基本假设。此外,SRSM 的各变量应满足非共线性,即 Kappa 值小于 100,但经过计算该模型的 Kappa 值高达 5570。从对模型的诊断结果看,SRSM 模型虽然预测效果较好,但应用于植被盖度的反演存在较大的缺陷。

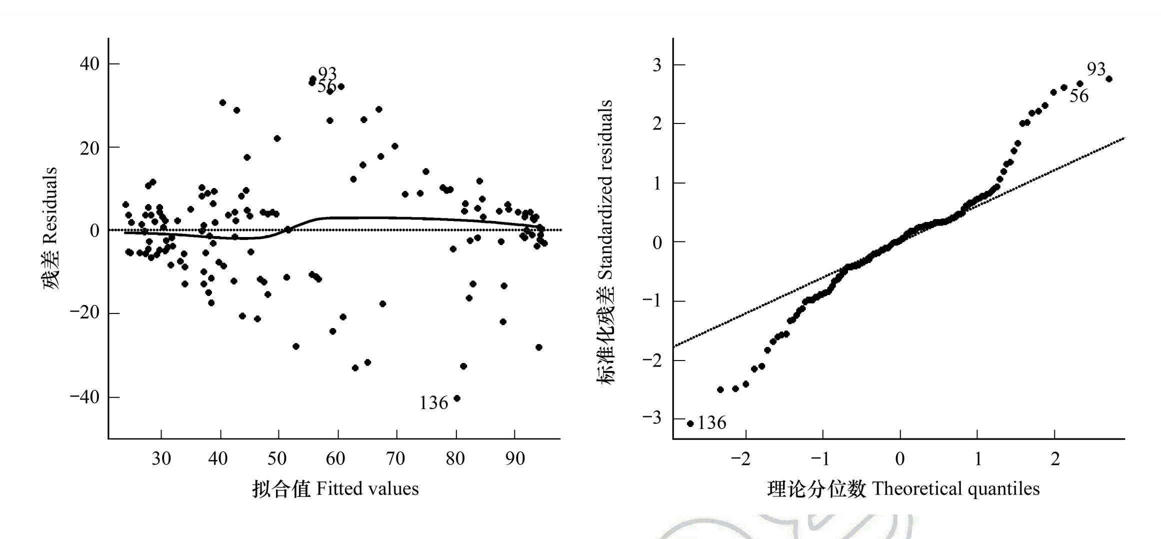


图 4 SRSM 模型的假设检验

Fig.4 Hypotheses testing of SRSM model

SRSM:基于 SR 数据 SAVI 及 MSAVI 指数的线性回归模型,Linear model based on SAVI and MSAVI from surface reflectance product

综上,本文认为随机森林模型相较于线性回归模型来说不但不受诸多假设条件和非共线性要求的制约,并能达到更好的预测效果,因此更适合作为植被盖度的反演模型。

4.3 2005—2015 年草场植被盖度变化趋势及不确定性分析

基于 L1、SR 以及 MODIS 数据分别得到的草场盖度平均值变化趋势如图 5 所示。除 2006 年之外,SR 和 L1 数据反演得到的植被盖度变化情况基本保持了一致的变化趋势,但 2005—2009 年,也即数据源为 ETM+的时间段内,基于 L1 的反演结果要明显小于 SR 数据的反演结果。SR 数据与不存在传感器差异问题的 MODIS 数据反演所得的植被盖度值相比,差异小于 5%,但 L1 数据的偏差较大,ETM+数据的预测结果普遍偏小,2006 年的植被盖度均值甚至偏小超过 20%。

为了分析传感器差异给反演结果造成的影响,本文进一步对比了同一时期内源自不同传感器光谱信息反演得到的植被盖度,对比结果如图 6 所示。该图显示了 2009 年 ETM+数据与 TM 数据,以及 2015—2016 年 ETM+数据与 OLI 数据分别反演得到的植被盖度的差异。

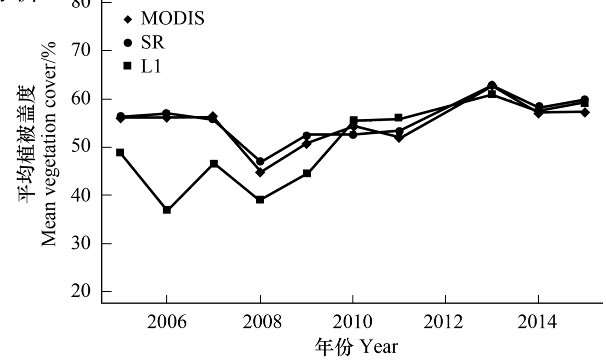


图 5 基于 MODIS、SR 及 L1 数据的抽样点盖度平均值

Fig.5 Mean vegetation cover of sampling points based on MODIS, SR and L1 data L1: 一级标准数据产品 Level1 Standard Data Product; SR: 经过校正和标准化处理的地表反射率数据产品 Surface Reflectance Data Products,

由图 6 的概率密度曲线可知,同一时段 L1 数据不同传感器对计算结果差异巨大,特别是 ETM+数据的预测结果普遍小于 TM 数据,上述结果与图 5 的结论一致。从具体计算的数值来看:对于 SR 数据,几乎 90% 的测试点位上不同传感器对计算结果产生的影响都能落在(-10%,11%)的区间之中;但对于 L1 数据,90%测试

chinaXiv:201804.01739v1

ChinaXiv合作期刊

点位的 ETM+与 OLI 的差异落在(-30%,11%)范围内,而 ETM+与 TM 的差异更是落在了(-40%,9%)之中。由此可见,不同传感器的差异在 L1 数据上表现的尤为突出,该数据不适于基于多元数据的草场植被盖度变化趋势研究,即便是经过校正以及标准化处理的 SR 数据,反演结果也存在约±10%的不确定性。

4.4 讨论

就反演方法而言,以往类似的研究多基于相关分析线性回归以及多项式模型[16-17,35-37],且模型的筛选通常仅仅通过 P 值以及决定系数 R^2 来评判,而不对模型预测结果做进一步验证。从本文的线性回归结果看,L1 数据和 SR 数据的一元线性模型都具有显著性且 R^2 水平相当,在以 MSAVI 作为解释变量时,L1 数据的拟合效果甚至要优于 SR 数据(表 3)。但从图 3 中可以清楚地看到前者的误差要远远大于后者。可见,P 值以及 R^2 并未帮助研究者对比出 SR 数据的优越性。更为重要的是,上述模型都属于参数估计类,若不对其基本假设进行验证,就无法保证所得结论的可靠性。本研究显示,所有的线性模型都未能通过方差齐性和残差正态性检验。而遗憾的是,迄今大多数研究都忽略了这种统计学的基本要求。

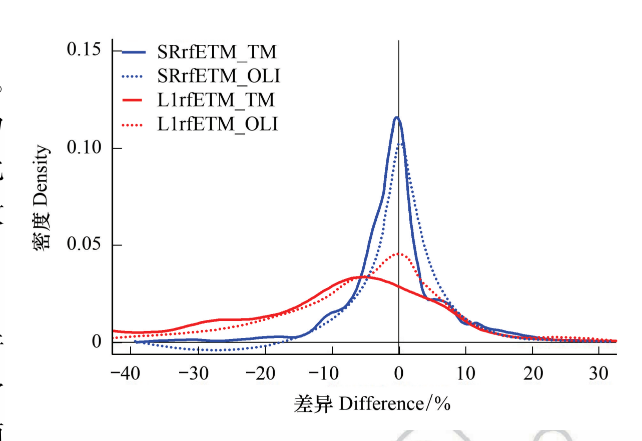


图 6 同时段不同数据源的预测结果差异

Fig.6 Difference in predicted values based on different dada sources in the same period

SRrfETM_TM:SR 数据基于 ETM+传感器与 TM 传感器随机森林反演结果的差异;SRrfETM_OLI:SR 数据基于 ETM+传感器与 OLI 传感器随机森林反演结果的差异;L1rfETM_TM:L1 数据基于 ETM+传感器与 TM 传感器随机森林反演结果的差异;L1rfETM_OLI:L1 数据基于 ETM+传感器与 OLI 传感器随机森林反演结果的差异

需要强调的是,除了模型本身之外,数据源也是影响预测效果的重要因素。由图 5 中基于 MODIS 数据反演得到的变化趋势可知,近 10 年来草场盖度的平均值总的变化幅度不足 20%左右,而 2006 年 L1 数据反演所得之盖度平均值与 MODIS 数据计算结果的偏差就已经超过了 20%。该问题在图 6 中体现得更加突出。这便意味着,传感器差异带来的影响甚至可能超过草场变化本身。虽然本研究所用的 SR 数据与传统研究常用的 L1 数据相比,已经在一定程度是减少了这类影响[38],但仍未将其完全消除。因此,在涉及长时间序列草场植被变化以及驱动力的研究中,对这种影响进行量化就显得尤为必要。

本文的局限性主要体现在以下几个方面:首先,使用的实地监测数据为 1m×1m 样方的测量值,而遥感影像为 30m×30m 精度,如果监测点周围数 10m 范围内不能保持均一性,拟合结果会因此受到影响;其次,本研究所涉及的草场种类众多,但植被指数计算涉及的参数均采用了 USGS 网站提供产品指南中的推荐值,后续研究中应对不同类型草场分别设定参数,以提高结果的可靠程度;再次,本研究初步估算了 2005—2015 年草场盖度的平均值,今后可进一步应用随机森林模型分析草场植被盖度的空间变化特征;最后,本文仅初步研究了传感器差异给反演结果造成的影响,如何通过归一化处理等方法减小这种差异将是未来研究中需要探索的问题。

5 结论

1) 本文通过随机森林回归方法建立了基于 Landsat 系列卫星数据的草场植被盖度反演模型,并应用该模型分析了近 10 年来布尔津县草场盖度的变化趋势。

2) 通过与线性回归反演模型的对比可知,随机森林回归方法不但能规避线性回归苛刻的假设条件,同时可超越绝大多数线性回归模型的预测能力。

3) 模型不确定性方面,Landsat ETM+的标准数据反演得到的植被盖度较之 TM 和 OLI 数据普遍偏小;地表反射率数据虽然可以大幅降低 Landsat 系列不同传感器对反演结果的影响,但反演得到的植被盖度仍存在

chinaXiv:201804.01739v1

www.ecologica.cn

ChinaXiv合作期刊

(-10%,11%)的不确定性。

致谢:北京大学生命科学学院王昊老师提供 MODIS 数据,中国环境科学研究院张玉波博士和付刚在遥感影像处理方面给予帮助,大理大学东喜马拉雅资源与环境研究所任国鹏老师为本研究提出建议,特此致谢。

参考文献(References):

[1] 樊江文, 钟华平, 陈立波, 张文彦. 我国北方干旱和半干旱区草地退化的若干科学问题. 中国草地学报, 2007, 29(5): 95-101.

[2] 黄敬峰, 桑长青. 天山北坡中段天然草场牧草产量遥感动态监测模式. 干旱区资源与环境, 1993, 7(2): 53-59.

[3] 李京, 陈晋, 袁清. 应用 NOAA/AVHRR 遥感资料对大面积草场进行产草量定量估算的方法研究. 自然资源学报, 1994, 9(4): 365-368.

[4] Hmimina G, Dufrêne E, Pontailler J Y, Delpierre N, Aubinet M, Caquet B, De Grandcourt A, Burban B, Flechard C, Granier A, Gross P, Heinesch B, Longdoz B, Moureaux C, Ourcival J M, Rambal S, Saint André L, Soudani K. Evaluation of the potential of MODIS satellite data to predict vegetation phenology in different biomes: an investigation using ground-based NDVI measurements. Remote Sensing of Environment, 2013, 132: 145-158.

[5] Geerken R, Batikha N, Celis D, DePauw E. Differentiation of rangeland vegetation and assessment of its status: field investigations and MODIS and SPOT VEGETATION data analyses. International Journal of Remote Sensing, 2005, 26(20): 4499-4526.

[6] Chen S B, Rao P. Land degradation monitoring using multi-temporal Landsat TM/ETM data in a transition zone between grassland and cropland of northeast China. International Journal of Remote Sensing, 2008, 29(7): 2055-2073.

[7] Feng Y, Lu Q, Tokola T, Liu H, Wang X. Assessment of grassland degradation in Guinan county, Qinghai Province, China, in the past 30 years. Land Degradation & Development, 2009, 20(1): 55-68.

[8] Geerken R, Zaitchik B, Evans J P. Classifying rangeland vegetation type and coverage from NDVI time series using Fourier Filtered Cycle Similarity. International Journal of Remote Sensing, 2005, 26(24): 5535-5554.

[9] Cai H Y, Yang X H, Xu X L. Human-induced grassland degradation/restoration in the central Tibetan Plateau: the effects of ecological protection and restoration projects. Ecological Engineering, 2015, 83: 112-119.

[10] Wessels K J, Prince S D, Carroll M, Malherbe J. Relevance of rangeland degradation in semiarid northeastern South Africa to the nonequilibrium theory. Ecological Applications, 2007, 17(3): 815-827.

[11] 魏小琴. 阿勒泰地区 NDVI 变化及其主要驱动因子分析[D]. 乌鲁木齐: 新疆农业大学, 2015.

[12] Röder A, Udelhoven T, Hill J, Del Barrio G, Tsiourlis G. Trend analysis of Landsat-TM and -ETM+ imagery to monitor grazing impact in a rangeland ecosystem in Northern Greece. Remote Sensing of Environment, 2008, 112(6): 2863-2875.

[13] Li J Y, Yang X C, Jin Y X, Yang Z, Huang W G, Zhao L N, Gao T, Yu H D, Ma H L, Qin Z H, Xu B. Monitoring and analysis of grassland desertification dynamics using Landsat images in Ningxia, China. Remote Sensing of Environment, 2013, 138: 19-26.

[14] 杨强, 王婷婷, 陈昊, 王运动. 基于 MODIS EVI 数据的锡林郭勒盟植被覆盖度变化特征. 农业工程学报, 2015, 31(22): 191-198.

[15] 吕长春, 王忠武, 钱少猛. 混合像元分解模型综述. 遥感信息, 2003, (3): 55-58, 60-60.

[16] 郭辉, 黄粤, 李向义, 包安明, 宋洋, 孟凡浩. 基于多尺度遥感数据的塔里木河干流地区植被覆盖动态. 中国沙漠, 2016, 36(5): 1472-1480.

[17] 马中刚, 孙华, 王广兴, 林辉, 余宇晨, 邹琪. 基于 Landsat 8-OLI 的荒漠化地区植被覆盖度反演模型研究. 中南林业科技大学学报, 2016, 36(9): 12-18.

[18] 李欣海. 随机森林模型在分类与回归分析中的应用. 应用昆虫学报, 2013, 50(4): 1190-1197.

[19] Cutler D R, Edwards Jr T C, Beard K H, Cutler A, Hess K T, Gibson J, Lawler J J. Random forests for classification in ecology. Ecology, 2007, 88(11): 2783-2792.

[20] 金宇, 周可新, 方颖, 刘欣. 基于随机森林模型预估气候变化对动物物种潜在生境的影响. 生态与农村环境学报, 2014, 30(4): 416-422.

[21] 王盼, 陆宝宏, 张瀚文, 张巍, 孙银凤, 季妤. 基于随机森林模型的需水预测模型及其应用. 水资源保护, 2014, 30(1): 34-37, 89-89.

[22] Rodríguez-Veiga P, Wheeler J, Louis V, Tansey K, Balzter H. Quantifying forest biomass carbon stocks from space. Current Forestry Reports, 2017, 3(1): 1-18.

[23] Mansour K, Mutanga O. Classifying increaser species as an indicator of different levels of rangeland degradation using WorldView-2 imagery. Journal of Applied Remote Sensing, 2012, 6(1): 063558.

[24] Chander G, Markham B L, Helder D L. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sensing of Environment, 2009, 113(5): 893-903.

[25] Vogelmann J E, Gallant A L, Shi H, Zhu Z. Perspectives on monitoring gradual change across the continuity of Landsat sensors using time-series

chinaXiv:201804.01739v1

www.ecologica.cn

ChinaXiv合作期刊

data. Remote Sensing of Environment, 2016, 185: 258-270.

[26] Pravalie R, Sîrodoev I, Peptenatu D. Detecting climate change effects on forest ecosystems in Southwestern Romania using Landsat TM NDVI data. Journal of Geographical Sciences, 2014, 24(5): 815-832.

[27] Pattison R R, Jorgenson J C, Raynolds M K, Welker J M. Trends in NDVI and tundra community composition in the arctic of NE Alaska between 1984 and 2009. Ecosystems, 2015, 18(4): 707-719.

[28] 毛志春, 宋宇, 李蒙蒙. 基于 MODIS 反演数据的河套地区荒漠化研究. 北京大学学报: 自然科学版, 2015, 51(6): 1102-1110.

[29] Liu B, You G Y, Li R, Shen W S, Yue Y M, Lin N F. Spectral characteristics of alpine grassland and their changes responding to grassland degradation on the Tibetan Plateau. Environmental Earth Sciences, 2015, 74(3): 2115-2123.

[30] USGS. Landsat surface reflectance-derived spectral indices, Version 3.3. 2016. [2016-12-13] https://landsat.usgs.gov/sites/default/files/documents/si_product_guide.pdf

[31] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.

[32] Liaw A, Wiener M. Classification and regression by randomForest. R News, 2002, 2(3): 18-22.

[33] Echeverry-Galvis M A, Peterson J K, Sulo-Caceres R. The social nestwork: tree structure determines nest placement in Kenyan weaverbird colonies. PLoS One, 2014, 9(2): e88761.

[34] Biau G, Scornet E. A random forest guided tour. TEST, 2016, 25(2): 197-227.

[35] Brinkmann K, Dickhoefer U, Schlecht E, Buerkert A. Quantification of aboveground rangeland productivity and anthropogenic degradation on the Arabian Peninsula using Landsat imagery and field inventory data. Remote Sensing of Environment, 2011, 115(2): 465-474.

[36] 万红梅, 李霞, 董道瑞. 基于多源遥感数据的荒漠植被覆盖度估测. 应用生态学报, 2012, 23(12): 3331-3337.

[37] 董建军, 牛建明, 张庆, 康萨如拉, 韩芳. 基于多源卫星数据的典型草原遥感估产研究. 中国草地学报, 2013, 35(6): 64-69.

[38] Claverie M, Vermote E F, Franch B, Masek J G. Evaluation of the Landsat-5 TM and Landsat-7 ETM+ surface reflectance products. Remote Sensing of Environment, 2015, 169: 390-403.

chinaXiv:201804.01739v1

www.ecologica.cn